Ecological Mining - A Case Study on Dam Water Quality

Manuel Filipe Santos[1], Paulo Cortez[1], Hélder Quintela[1], José Neves[2], Henrique Vicente[3] & José Arteiro[3]

[1] *Department of Information Systems, University of Minho, Portugal*
[2] *Department of Computer Science, University of Minho, Portugal*
[3] *Department of Chemistry, University of Évora, Portugal*

Abstract

The automatic assessment of barrage water quality is very restricted due to the distances, the number of biochemical parameters to be considered and the financial resources spent to obtain their values. To this scenario it should be added the latency times between the sampling moment and the outcome of the laboratory analyses.

Although the idea of considering sensors for remote acquisition of data is not new, there are some constraints to be addressed, like: the existence of sensors to measure the pertinent parameters and their efficiency, the costs involved and the possibility of remote sensing.

The application of this alternative is highly dependent on the relevance of the candidate parameters. At this point, the Data Mining (DM) approach assumes an important role, in the sense it can reveal the relative importance of the parameters, as well the prediction models to determine the water quality and finally the associated accuracies.

This paper introduces a decision framework to support the selection of

biochemical parameters to be considered in remote sensing of water contained in barrages. The framework enables the comparison of the efficiency of two kinds of models, using decision trees. The first one uses all the water quality indicators, including the time and cost consuming variables, while the second model is based only on remotely real-time acquired parameters. When comparing both strategies under several criteria (e.g., cost, time and confidence), the latter method showed to be best alternative.

Keywords: Data Mining, Knowledge Discovery from Databases, Decision Support, Water Quality, Decision Trees.

1 Introduction

The interest on ecological mining has been growing in last decades. Relationships between living communities and their abiotic environment can be highly nonlinear and ecological models have to reflect this to be realistic [1]. Indeed, several machine learning algorithms have been used to find patterns in river water quality databases, such as Artificial Neural Networks and Decision Trees [2][3][4]. While ANNs have been used extensively in ecological modelling [5] [6], Decision Trees have the advantage of expressing regularities explicitly and thus being easy to inspect for ecological validity.

Currently, the assessment of dam water quality is done through analytical methods, which is very restricted approach due to the distances, the number of parameters to be considered and the financial resources spent to obtain their values. Moreover, to this context, it should be added the latency times between the sampling moment and the outcome of the laboratory analyses. Due to these constraints, the development of Data Mining (DM) based models [7] in conjunction with the development of a Decision Support System [8], is a better alternative for the quality management of water resources.

In this paper, it is exploited an approach to make the assessment of water quality easier, cheaper and faster by DM models, taking in account the parameters that could be measured in real time by automatic mechanisms (e.g., sensors) and accessed through a communications infrastructure. This simplified model is compared with a more complex one, which uses a high number of parameters.

The paper is organized as follows: first, the ecological data is presented and described; then, the decision trees are introduced; next, the experiments performed are described, being the results analysed in terms of several criteria; finally, closing conclusions are drawn.

2 Materials and Methods

2.1 Ecological Data

The data used in this study was collected from 1982 to 2003, in three Large Portuguese Dams [9] (Table 1), located in the High Alentejo region of Portugal, containing a total of 998 records with 170 chemical, physical, and microbiological parameters. These are considered the main attributes that may reflect the water quality at a particular point in time. Table 2 shows a synopsis of some relevant water quality input features.

Table 1: The three Portuguese Dams considered in this study

	Dam		
Characteristics	*Divôr*	*Monte Novo*	*Vigia*
Location	Évora	Évora	Évora
River	Divor	Degebe	Vale do Vasco
Bassin	Tagus	Guadiana	Guadiana
Purpose	Irrigation Water Supply	Irrigation Water Supply	Irrigation Water Supply
Volume (m3)	255 000	31 000	284 000

The classification of the quality of superficial water mass can be done by two evaluation criteria [10]: the first one classifies waters in terms of treatment for human consuming; the second classifies water masses taking in account characteristics for multiple uses. Since the later criterion allows a wider use and it is also adopted by INAG, the Portuguese water management service, it will be adopted in this work (Table 6, Appendix A). Therefore, water will be classified in the non linear scale **A**, **B**, **C**, **D**, or **E** [11][12], where **A** denotes no pollution and **E** denotes extreme pollution, which represents serious risks in terms of public and environmental health (Figure 1). The original dataset presented biased distributions: in 57.8% of the observations the water quality of the dam is very polluted (**D**); 34.5% is polluted water (**C**); 6.1% is weak polluted (**B**); 1.7% is extremely polluted (**E**); and no non polluted (**A**) cases are found.

Before attempting the DM modelling, the data was pre-processed. The original dataset contained attributes with missing values. In particular, the parameter chlorophyll presented a high number of blank values (202). Since it was not possible to obtain the correct values the blank registers were discarded [13],

remaining a total of 722 examples. In addition, in order to enhance the Decision Tree learning, the chlorophyll indicator was levelled from the continuous interval [0.0,…,100.0] to the discrete domain {1,2,3,4,5} [13].

Table 2: The main quality indicators of water

Monthly acquired	**Bi-monthly acquired**	**Yearly acquired**
pH value	Dissolved Iron	Fluoride
Colour	Manganese	Boron
Total Suspended Solids	Cooper	Arsenic
Temperature	Zinc	Cadmium
Conductivity	Sulphate	Total Chromium
Odour	Surfactants	Lead
Nitrate	Phenols	Selenium
Chloride	Azote	Mercury
Total Reactive Phosphorus	Kjeldahl Nitrogen	Barium
Chemical Oxygen demands	Faecal Streptococcus	Cyanide
Dissolved Oxygen	-	Dissolved Hydrocarbons
5-Day Biochemical Oxygen Demands	-	Polynuclear Aromatic Hydrocarbons
Ammonia Nitrogen	-	Total Pesticides

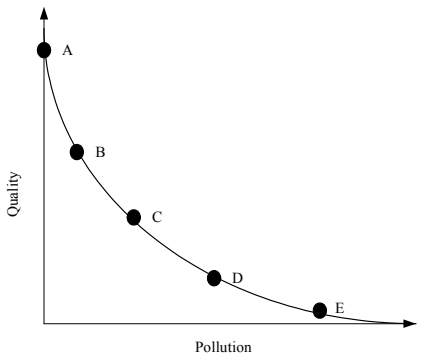

Figure 1: The water quality classes vs. the pollution factor

Finally, the non monthly parameters of Table 2 were transformed into monthly ones, by replicating the last known value. For instance, if the last value of a bi-monthly variable was acquired in May, then the same value will be used in June.

2.2 Decision Trees

The prediction of river water quality, according to the criteria followed by INAG, was defined as a classification problem. The Decision Tree is one of the most efficient and popular DM classification algorithms. It adopts a branching structure of nodes and leaves, where the knowledge is hierarchically organized. Each node tests the value of some feature, while each leave assigns a class label. In this study, the C5.0 algorithm [14] was used to induce the decision trees, under the SPSS Clementine System [15]. The use of decision trees enables the automatic extraction of production rules that can be incorporated in a Decision Support System, using for example, Structured Query Language (SQL) commands or using the Predictive Markup Language (PMML) specification [16].

3 Results

3.1 Framework

Attending that the water quality analysis is very difficult and time consuming, it was decided to develop the experimentation with two different strategies. The first approach (**Model 1**) is based in a Decision Tree built using all the water quality indicators, while the second (**Model 2**) uses only the parameters that can be measured by sensors in real-time. These two approaches will be compared in a framework, under the following criteria: time to get the results; cost of the analysis; acquisition in site of the dam; real-time acquisition; predictive accuracy; and real time diagnosis of water quality.

3.2 Tests

The classification models for water quality were developed using the C5.0 algorithm. To insure statistical significance of the attained results, 10 runs were applied in all tests, being the accuracy estimates achieved using the Holdout method [17]. In each simulation, the available data is randomly divided into two mutually exclusive partitions: the training set, with 2/3 of the available data and

used during the modelling phase; and the test set, with the remaining 1/3 examples, being used after training, in order to compute the accuracy values.
A common tool for classification analysis is the confusion matrix [18], a matrix of size L x L, where L denotes the number of possible classes. This matrix is created by matching the predicted (test result) and actual (water quality real condition) values. Since no **A** and **E** cases are in the dataset (the E cases were removed due to the presence of missing values), L was set to 3 (classes **B**, **C** and **D**).
In preliminary experiments, several input feature selections were tested for **Model 2**. This task was guided by an expert in the field of environment science. After this procedure, a total of 7 real-time acquired variables were selected: oxygen, pH value, transparency, chloride, wind speed, precipitation and temperature.
Table 3 shows examples of the decision rules obtained for each strategy, while Figure 2 displays the specification of Model 1, according to the PMML v.2.1 standard.

```
<?xml version="1.0" encoding="UTF-8" ?>
<PMML version="2.1" xmlns="http://www.dmg.org/PMML-2_1"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.dmg.org/PMML-2_1 pmml-2-1.xsd">
<Header copyright="Copyright (c) 2002 Integral Solutions Ltd., All Rights Reserved.">
<Application name="Clementine" version="8.1" />
</Header>
<DataDictionary numberOfFields="14">
<TreeModel modelName="Water_Quality" functionName="classification" algorithmName="RuleSet">
...
<Node score="3" recordCount="21">
  <Extension name="x-nrCorrect" value="22" extender="spss" />
  <Extension name="x-nodeId" value="17" extender="spss" />
  <CompoundPredicate booleanOperator="and">
        <SimplePredicate field="oxygen" operator="greaterThan" value="61.799999" />
        <SimplePredicate field="chemical oxygen demand" operator="greaterThan" value="19.799999" />
        <SimplePredicate field="transparency" operator="greaterThan" value="0.3" />
        <SimplePredicate field="ammonia" operator="greaterThan" value="0.002" />
        <SimplePredicate field="ammonia" operator="lessOrEqual" value="0.009" />
  </CompoundPredicate>
  <ScoreDistribution value="1" recordCount="0" />
  <ScoreDistribution value="2" recordCount="0" />
  <ScoreDistribution value="3" recordCount="22" />
  <ScoreDistribution value="4" recordCount="0" />
</Node>
...
</TreeModel>
</PMML>
```

Figure 2: An extract of PMML specification for Model 1

Table 3: A snapshot of the Decision Rules for each model

Model 1	Model 2
…	…
If oxygen > 61,8 and chemical oxygen demands > 19,8 and transparency> 0,3 and ammonia > 0,002 and ammonia ≤ 0,009 Then → C (n=**21**, confidence=**1,0**)	If oxygen ≤ 61,8 and pH value > 7,17 and transparency ≤ 1,3 and chloride ≤ 45 Then → D (n=**29**, confidence= **0,895**)
…	…

3.3 Discussion

Table 4 presents the confusion matrixes for each approach, where the values denote the average of the 10 runs. The results reveal that the second model is more accurate in predicting much polluted cases (**D**), with an accuracy of 96.8%. Yet, the first model outperforms the latter when predicting other classes. **Model 1** presents an accuracy of 77%, while Model 2 denotes a 74% classification rate. Although this comparison considers all 3 the classes, the test sets only contain a very limited number of weak polluted (**B**) examples. Thus, it makes sense to perform an analysis considering only **D** and **C** classes, where the other values are transformed into the nearest class. Under this setting, the scenario changes, with the second model outperforming the former one (83.1% vs. 78.5%).

Table 4: The confusion matrix for each model

Model 1

Class	Training Set			Test Set		
	D	**C**	**B**	**D**	**C**	**B**
D	66	3	0	25	6	0
C	2	58	0	8	25	1
B	0	0	10	0	0	3

Model 2

Class	Training Set			Test Set		
	D	**C**	**B**	**D**	**C**	**B**
D	68	1	0	30	1	0
C	11	47	2	10	21	3
B	1	3	6	2	1	0

Besides a good predictive accuracy, there are other important factors for model selection (Table 3), such as the necessity to get the results faster. When considering the other criteria, **Model 2** is clearly the best choice.

Table 5: The framework for the model assessment

Evaluation Criteria	**Model 1**	**Model 2**
Time to obtain analytical results	Time-consuming	Fast
Cost of the analysis	Expensive	Reasonable
Acquisition of the value of the parameters in site	No	Yes
Acquisition of the value of the parameters remotely	No	Yes
Real time acquisition of the value of the parameters remotely	No	Yes
Predictive Accuracy	76.8%/78.5%	73.9%/83.1%
Water Quality in real time	No	Yes

4. Conclusion and Further Work

The use of DM techniques can solve complex problems in environmental applications, as the real-time diagnosis of water quality in dam lagoons. In this work, two classification models were tested, using decision trees. The first adopted 177 input variables while the latter only considered real-time acquired data. The experiments were conducted in order to test several input feature selection configurations, leading to a simpler decision tree based on only 7 variables (**Model 2**). The results so far obtained give an overall accuracy of 77/79% for **Model 1** depending if 3 or 2 classes are predicted. On the other hand, **Model 2** presented a classification rate of 74% (for 3 classes) and 83% (for 2 classes). When the analysis is performed under additional criteria (e.g., time and costs), the latter approach clearly excels the former. The obtained decision trees, which are easy to interpret, were validated by experts. The proposed approach opens room for the development of automatic tools for environmental decision support, which are expected to enhance the ecological response. Indeed, in future work, it is intended to apply these techniques in real environments, in a on-line learning, where dam sensors feed directly data into a decision support system.

References

[1] Dzeroski, S., "Applications of symbolic machine learning to ecological modelling," Ecological Modelling, pp. 263-273, 2001.

[2] Walley, W. J., Hawkes, H.A., Boyd, M., "Application of Bayesian inference to river water quality surveillance.," presented at Applications of Artificial Intelligence in Engineering Computational Mechanichs Publications., Southampton, England, 1992.

[3] Dzeroski, S., Grbovic, J., "Knowledge Discovery in a water quality database," presented at First International Conference on Knowledge Discovery and Data Mining, Menlo, Park, CA, 1995.

[4] Walley, W.J., Dzeroski, S., Biological monitoring: A comparison between Bayesian, neural and machine learning methods of water quality classification, International Symposium on Environmental Software Systems, 1996.

[5] Maier, H.R., Dandy, G.C., Burch, M.D., "Use of artificial neural networks for moddeling cyanobacteria Anabaena spp. in the River Murray, South Australia," Ecological Modelling, pp. 257-272, 1998.

[6] Schleiter, I.B., Borchardt, D., Wagner, R., Dapper, T., Schmidt, K.-D., Schmidt, H.-H, Werner, H., "Modelling water quality, bioindication and population dynamics in lotic ecosystems using neural networks.," Ecological Modelling, vol. 120, pp. 271-286, 1999.

[7] Fayyad, U.M., Piatetski, G.S., Smith, P., Advances in Knowledge Discovery and Data Mining, The MIT Press, Massachussets, USA, 1996.

[8] Turbam, E., Aronson, J. E., Decision Support Systems and Intelligent Systems, Prentice Hall.

[9] Large Dams in Portugal, LNEC, http://www-ext.lnec.pt/lGb/index.phtml.

[10] Vicente, H., Especificação e Prototipação de Sistemas de Gestão e Controlo da Qualidade da Água de Albufeiras, PhD Thesis, Évora, Portugal, 2004.

[11] Direcção Regional do Ambiente e Ordenamento do Território - Alentejo, Anuário de recursos hídricos do Alentejo – Ano Hidrológico 2001/2002, Portugal, 2003.

[12] Matoso, A., Rasga, M., Santana, M., Murteira, M., Principais Albufeiras do Alentejo monitorizadas, Comissão de Coordenação da Região do Alentejo – Direcção de Serviços de Monitorização Ambiental, Portugal, 2004.

[13] Pyle, D., "Data preparation for Data Mining," 1999.

[14] Quilan, J.R., Bagging Boosting and C4.5, Proceedings of Fourteenth National Conference on Artificial Intelligence.

[15] SPSS Inc., http://www.spss.com.

[16] Data Mining Group, http://www.dmg.org/pmml-v2-1.html.

[17] Souza, J., Matwin, S., Japkowicz, N., Evaluating Data Mining Models: A Pattern Language, Proceedings of the 9th Conference on Pattern Language of Programs, Illinois, USA, 2002.

[18] Kohavi, R., Provost, F., Glossary of Terms, Machine Learning, 30(2/3), pp. 271-274, 1998.

Appendix A

Table 6: Ranges proposed by INAG to classify the quality of superficial water

Parameter	Categories/Classe				
	A	B	C	D	E
pH value (Sorensen)	6,5 - 8,5		6,0 - 9,0		5,5 - 9,5
Temperature / ºC	≤ 20	> 20 ≤ 25	> 25 ≤ 28	> 28 ≤ 30	> 30
Condutivity / µS cm^{-2}	≤ 750	> 750 ≤ 1000	> 1000 ≤ 1500	> 1500 ≤ 3000	> 3000
Total Suspended Solids / mg dm^{-3}	≤ 25,0	> 25,0 ≤ 30,0	> 30,0 ≤ 40,0	> 40,0 ≤ 80,0	> 80,0
Dissolved Oxygen (% sat)	≥ 90	< 90 ≥ 70	< 70 ≥ 50	< 50 ≥ 30	< 30
Oxidability / $mg_{O_2}\,dm^{-3}$	≤ 3,0	> 3,0 ≤ 5,0	> 5,0 ≤ 10,0	> 10,0 ≤ 25,0	> 25,0
5-day BOD (20ºC) / $mg_{O_2}\,dm^{-3}$	≤ 3,0	> 3,0 ≤ 5,0	> 5,0 ≤ 8,0	> 8,0 ≤ 20,0	> 20,0
COD / $mg_{O_2}\,dm^{-3}$	≤ 10,0	> 10,0 ≤ 20,0	> 20,0 ≤ 40,0	> 40,0 ≤ 80,0	> 80,0
Ammonia nitrogen / $mg_{NH_4^+}\,dm^{-3}$	≤ 0,10	> 0,10 ≤ 1,00	> 1,00 ≤ 2,00	> 2,00 ≤ 5,00	> 5,00
Nitrate / $mg_{NO_3^-}\,dm^{-3}$	≤ 5,0	> 5,0 ≤ 25,0	> 25,0 ≤ 50,0	> 50,0 ≤ 80,0	> 80,0
kjeldahl nitrogen / $mg_N\,dm^{-3}$	≤ 0,50	> 0,50 ≤ 1,00	> 1,00 ≤ 2,00	> 2,00 ≤ 3,00	> 3,00
Total reactive phosphorus / / $mg_{P_2O_5}\,dm^{-3}$	≤ 0,54		> 0,54 ≤ 0,94		> 0,94
Total coliforms / nº / 100 cm^3	≤ 50	> 50 ≤ 5000	> 5000 ≤ 50000	> 50000	
Fecal coliforms / nº / 100 cm^3	≤ 20	> 20 ≤ 2000	> 2000 ≤ 20000	> 20000	